Multilevel rhythms in multimodal communication

This paper has been accepted for publication in the *Philosophical Transactions of the Royal Society B: Biological Sciences*. For citation: Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., Schaefer, R., Wiggins, G. (Accepted). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi: 10.1098/rstb.2020.0334

Title: Multilevel rhythms in multimodal communication

Article type: 'Review and perspective paper'
Special Issue: 'Synchrony and rhythm interaction: from the brain to behavioural ecology'

Authors: Wim Pouw[1,2], Shannon Proksch[3], Linda Drijvers[1,2]**, Marco Gamba[4]**, Judith Holler[1,2]**, Christopher Kello[3]**, Rebecca S. Schaefer[5,6]**, Geraint A. Wiggins[7]**
***equal contribution, alphabetical order*

1. Donders Institute for Cognition, Brain and Behaviour, Radboud University, Nijmegen, The Netherlands
2. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
3. Cognitive and Information Sciences, University of California, Merced, USA
4. Department of Life Sciences and Systems Biology, University of Turin, Italy
5. Health, Medical and Neuropsychology unit, Institute for Psychology, Leiden University, Leiden, The Netherlands
6. Academy for Creative and Performing Arts, Leiden University, Leiden, The Netherlands
7. Vrije Universiteit Brussel, Belgium & Queen Mary University of London, UK

Correspondence: Wim Pouw (w.pouw@psych.ru.nl)

Acknowledgements: We would like to thank the organizers of the Lorentz workshop 'Synchrony and rhythm interaction' for their leadership in the field. WP is supported by a Donders Fellowship and is financially supported by the Language in Interaction consortium project 'Communicative Alignment in Brain & Behavior' (CABB). LD is supported by a Minerva Fast Track Fellowship from the Max Planck Society. LD and JH are supported by the European Research Council (CoG grant #773079, awarded to JH).

Abstract

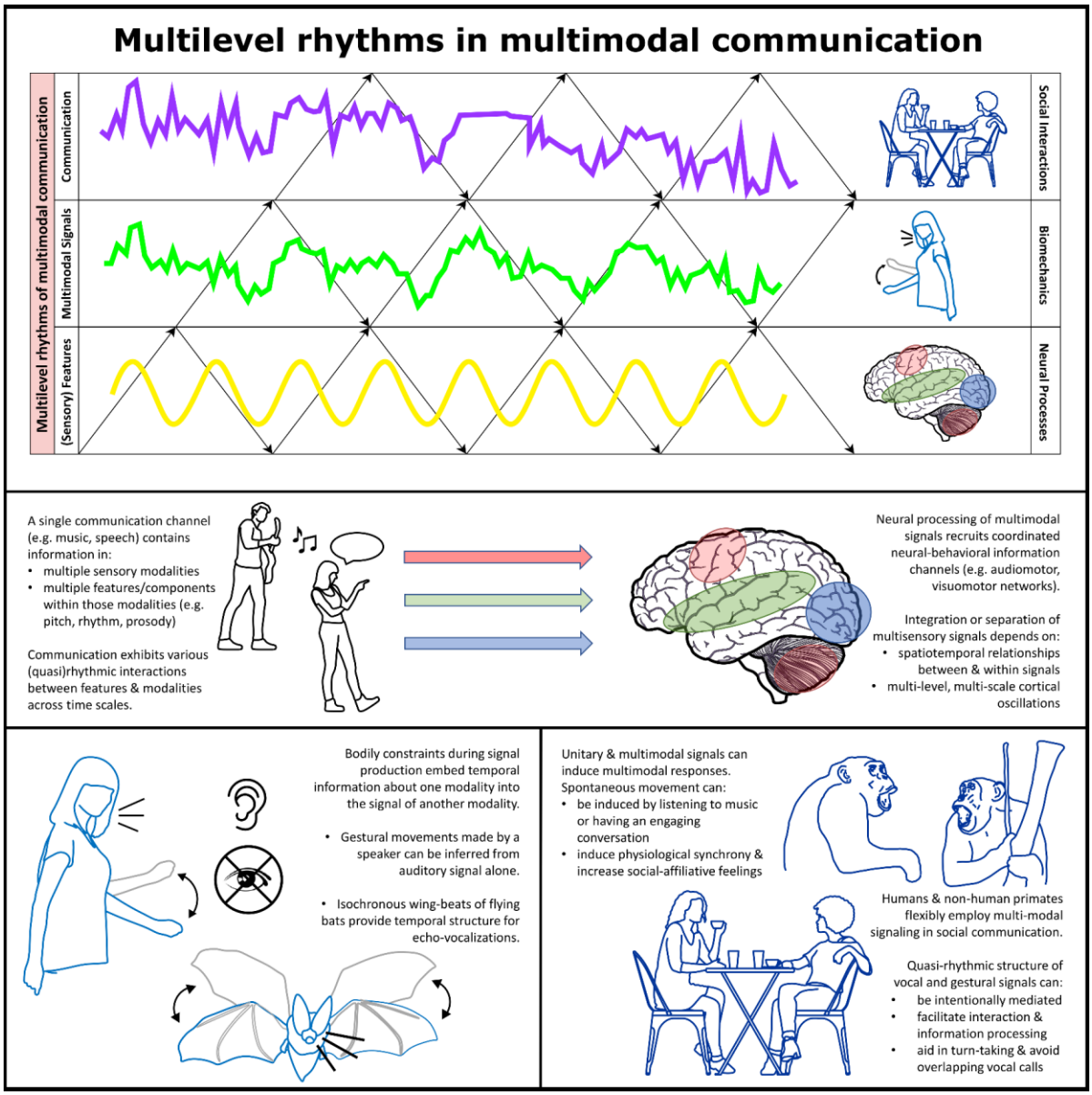
It is now widely accepted that the brunt of animal communication is conducted via several modalities, e.g. acoustic and visual, either simultaneously or sequentially. This is a laudable multimodal turn relative to traditional accounts of temporal aspects of animal communication which have focused on a single modality at a time. However, the fields that are currently contributing to the study of multimodal communication are highly varied, and still largely disconnected given their sole focus on a particular level of description or their particular concern with human or non-human animals. Here we provide an integrative overview of converging findings that show how multimodal processes occurring at neural, bodily, as well as social interactional levels each contribute uniquely to the complex rhythms that characterize communication in human and non-human animals. Though we address findings for each of these levels independently, we conclude that the most important challenge in this field is to identify how processes at these different levels connect.

Word count (in text, including references): **7068**
Keywords: Multimodal Communication, Rhythm, Multimodal Signaling, Cross-species, Interaction

Introduction

The rhythms animals can sustain in communicative perception and action characterize in great part their social ecological niche. It is only recently that disparate research fields have focused on the study of temporal aspects of communication as a truly multimodal process[1–3]. Lessons about the different scales or levels at which multimodal processes happen are however still scattered over different fields, such as psycholinguistics [3], neuroscience [4], and evolutionary biology [5]. The goal of this paper is to align some of the important findings of these fields concerning the different ways in which the brain, body, and social interaction each contribute uniquely to the temporal structure of multimodal communication (see Figure 1 for an overview). Although we overview findings at each level (neural, body, social) independently, we hope to stimulate investigation into potential interactions between levels. We provide some broad terminology for the phenomenon of multilevel rhythm of multimodal communication (section 1), and then overview rhythmic multimodal processes on the neural-cognitive (section 2), the peripheral body (section 3), and the social interactional level (section 4).

Figure 1. Multilevel rhythm in multimodal communication



Note. Graphical overview of how each level contributes uniquely to the rhythms sustained in multimodal communication. Figures are adapted from [6,7], and inspired by Gilbert Gottlieb's (1929-2006) view on epigenesis.

Section 1. Concepts and terminology

Multimodal processes interest researchers from largely disparate fields and consequently terminology varies [1,5,8], where related meanings potentially get lost in translation. In box 1 we have marked the terms and their senses that occur throughout our overview. This glossary also aims to capture a very general meaning of specialist terms offered to address a particular process in perception or production, or at a neural, structural, or behavioral level. The definitions are as general as possible, for instance so as to underline a continuity of the perception and production of multimodal signals or so as to include phenomena not traditionally treated as multimodal in nature. For example, in sign languages

both the hands as well as facial and labial expressions are combined in complex utterances [9]. Though such complex signs are designed to be received through one sensory channel and thus unimodal by common definitions (but see [10]). In our view signed languages are an example of a *multimodal production* in virtue of combining several otherwise independent signaling *modes*/signal *features*. Similarly, neural processes can be multimodal in our view, in virtue of coupling neural ensembles that independently would be tuned to differently structured information in the environment. Note, that we cannot address all the rich and varied (temporal) functions of complex multimodal signaling [5,8,10]. But in our review the common thread resonates with a recent overview by Halfwerk and colleagues (2019) who suggest that multimodal signaling functions are not exhausted by simply a) providing redundant backup information or b) combining multiple independent messages. Instead, what is central to temporal functioning of multimodal systems is that the resulting perception or production of a signal "is qualitatively different from the sum of the properties of its components" ([10], p. 2), i.e., has emergent properties [8].

Box 1 Definitions

General definition of phenomenon	Term	Context of term	Example
A distinct measurable aspect of a system, which can be measured independently of other aspects	Component	Ethology	Frequency or duration of a signal; an intellectual instance determining behavior
	Component	Mathematics; Electronic engineering (EE)	Partial at frequency x; Regions of energy concentration
	Feature	EE; Computer Science (CS);	Spectral centroid; Signal onset/offset; Duration of a signal
	Feature	Music **Current paper**	Pitch; Fundamental frequency; Rhythm; Harmony
Unitary communication event X which is informative about state of affairs Y to a receiver (1) and/or producer (2)	Cue (1)	Ethology	Size of an animal, not intentionally communicated
	Natural signs (1)	(Peircian) Semiotics	Footsteps in the sand, not intentionally communicated
	Sign (1 & 2)	(Peircian) Semiotics **Current paper**	Word or gesture, intentionally communicated; understood in a three place relation of sign, referential target, and the user of the sign
Sensory and/or effector communication channel conventionally treated as functionally separable from others	Modality	Neuroscience **Current paper**	Specific neural ensembles associated with processing of a specific sensory channel or structure
	Modality	Psycholinguistics; psychomusicology; Ethology **Current paper**	Audition; Vision; Touch (usually ascribed to senses of the receiver — the receiver processes light signals via the sense of vision)
	Mode	Movement science; **Current paper**	Whispering, phonating; In-phase, anti-phase synchrony; Resonance; Punching, kicking
A measurable aspect of a producing system, changing in time, which is used by a receiver system.	Signal	Mathematics; EE; CS; **Current paper**	Frequency, voltage, amplitude
	Signal	Ethology	A (sequence of) vocalization(s), or movement(s), etc *intentionally* produced for a receiver, e.g. a specific mating call

Informational, temporal, and/or mechanical coupling between two or more measurable aspects, the coupling of which benefits communicative purposes. The benefit can be for the producer (1) and/or the recipient (2)	Multimodal Cue *(1)*	Ethology **Current paper**	Information about body movement or size from vocal patterning; Indexical cues
	Multimodal signal *(1 & 2)*	Ethology; Psycholinguistics **Current paper**	Sonic communication with facial and/or manual gesture
	Multi-component signal *(1 and 2)*	Ethology	Combined vocal and visual signaling
	Coordination of modes *(1 and/or 2)*	Movement science; **Current paper**	Entrainment of neural ensembles for sensory integration; Coordination of respiratory, jaw, and articulatory modes for speaking; Gesture (person 1) and speech (person 2) interactions

Section 2. Neural level: Multimodal neural-cognitive processes

Here we present an overview of how temporal coupling in the production and perception of multimodal signals can be constrained by neural ensembles that are independently tuned towards specifically structured information in the environment. In their multimodal arrangement, they yield unique stabilities for tuning to the rhythms of multimodal communication. Furthermore, some neural ensembles are uniquely specialized to attune to multisensory information.

When integrating a cascade of sensory signals to form a unified, structured percept of the environment, the brain faces two challenges. First, integrating different sensory signals into a unified percept relies on solving the 'binding problem': whether signals need to be integrated or segregated. Second, these sensory signals require integration with prior and contextual knowledge to weigh their uncertainty.

The neural integration of multiple sensory signals is describable at several neural levels and measurable using wide-ranging methods (e.g., single unit recordings, optogenetics, EEG, MEG, fMRI, combined with psychophysical experiments [11–13]). Although the potential multisensory integration mechanisms are debated, the integration likelihood of two signals seems highly dependent on the degree of spatiotemporal coherence between those signals: unisensory signals that are closer in time and space have a greater likelihood of being integrated (cf. [3] and section 4). Both human and non-human animal research demonstrates that multisensory neurons in the superior colliculus respond more robustly to spatiotemporally congruent audiovisual cues than to individual sensory cues [14–16]. For example, in macaques (*Macaca mulatta*) single-unit activity measurements in one specific area in the superior temporal sulcus (anterior fundus) show unique sensitivity to facial displays when temporally aligned with vocal information, while other areas (anterior medial) are sensitive to facial displays alone [17]. Behavioral evidence of multisensory integration is shown in the territorial behavior of dart-poison frogs (*Epipedobates femoralis*), who aggress conspecifics more when auditory and visual cues are sufficiently spatiotemporally aligned [18]. Note though, multimodal temporal alignment need not entail synchronization but can specifically involve structured sequencing (i.e., alignment at a lag). This is evidenced by research on a taxa of flycatcher bird species (*Monarcha castaneiventris*) who are uniquely responsive to long-range-emitted song followed by seeing plumage color of potential territorial rivals as opposed to their, reversely ordered, synchronized, or unimodal presentation[19]. Integration by temporally aligned presentation can be a developmentally acquired disposition,

as research in cats shows that development of multisensory integration in the superior colliculus is dependent on exposure to spatiotemporally coherent visual and auditory stimuli early in life [12].

Although lower-level and higher-level multimodal integration mechanisms are not well understood, both feedback and feedforward interactions between early and higher-level cortices might be relevant for integration. Specifically, it has been hypothesized that synchronized neural oscillations provide a mechanism for multisensory binding and selecting information that matches across sensory signals [20]. Here, coherent oscillatory signals are thought to allow for functional connectivity between spatially distributed neuronal populations, where low-frequency neural oscillations provide temporal windows for cross-modal influences [21]. This synchronization can occur through neural entrainment and/or phase resetting, which might be relevant for phase reorganization of ongoing oscillatory activity, so that high-excitability phases align to the timing of relevant events [21]. New methods, such as rapid invisible frequency tagging [22–24], might clarify how multisensory signals are neurally integrated, and what the role of low-frequency oscillations is in this process over time. Moreover, novel approaches focusing on moment-to-moment fluctuations in oscillatory activity combined with methods with increased spatial resolution (e.g., ECoG/depth-electrode recordings), could significantly advance our knowledge of the role of oscillatory activity in routing and integrating multisensory information across different neural networks [21]. This will be especially relevant in more complex, higher-level multimodal binding scenarios, such as (human) communication.

Communicative signals in naturalistic settings arguably include multiple features that work together to maximize their effectiveness. Different sensory modalities may operate at different timescales, with specific well-matched combinations of features across modalities, leading to common cross-modal mappings that are intuitively associated (e.g. visual size and auditory loudness, cf. [25–27]). Prominent well-matched cross-modal mappings (see sections 3 and 4) are sensorimotor mappings: signals transmitted to, from, or within visuomotor and auditory-motor systems. Given the high sensitivity of the auditory system for periodic signals aligned with motor periodicities [28,29], auditory signals often entrain movement, with examples seen in various kinds of joint action (e.g., marching or other timed actions). Less commonly, visual signals serve this purpose, as seen in musical conductors. Moreover, perception of both auditory and visual rhythms shares neural substrates with the motor system in terms of timing mechanisms [30]. While the auditory- versus visual modality seems better suited to guide movement [31], it appears that within different sensory modalities, different features may be better suited to cue movement [32]. For example, movement is most easily cued by discrete events in the auditory domain (e.g., beeps), followed by continuously moving objects in the visual domain (e.g. moving bars)[32]. For discrete visual stimuli (e.g., flashes), or continuous auditory stimuli, (e.g., a siren), sensorimotor synchronization is less stable (see for similar results in audiovisual speech: [33]). In contrast to humans, Rhesus macaques (*Macaca mulatta*) more easily synchronize to discrete visual cues [34] perhaps due to weaker audiomotor connections in the Macaque brain [35]. These findings indicate that multimodal perception is not simply a matter of adding more modalities, but rather the combination of temporal structure and signal content, affecting behavioral performance and neural activations [36,37]. Moreover, compelling arguments based on multimodal mating signals in a range of species as as reviewed by Halfwerk and colleagues [10] suggests that

exactly this integration of signals, leading to a multimodal percept rather than a main and a secondary modality, is what makes them informative.

Behavioral and neural studies show that temporal structures in one sensory domain can affect processing in another. Examples are auditory [38] or even multisensory rhythmic cues such as music or a metronome [28] not only regularizing movement (i.e. changing motion trajectories as compared to uncued movements), but also entraining visual attention [37], by increasing visual sensitivity at time points predicted to be salient by an auditory stimulus. The neural underpinnings of such interactions are largely unclear. Music-cued versus non-cued movement leads to additional neural activation in motor areas, specifically cerebellum [41,42], suggesting that the neural activations related to multimodal processing are synergetic. This may explain findings of enhanced learning with multimodal cues, for instance when auditory feedback of movement (or sonification) is provided [43,44]. Even when multimodal embedding of motor learning does not show clear behavioral increases, differences in learning-related neural plasticity were reported for novices learning a new motor sequence to music as compared to without [45], suggesting that the learning process is implemented qualitatively differently [46].

Taken together, different sensory modalities, and the features embedded in these signals, have different sensitivities for specific timescales, making some features especially suitable for cross-modal combinations. When investigating features that naturally combine, behavioral and neural responses emerge which amount to more than a simple addition of multiple processes.

Section 3. Body level: Multimodal signaling and peripheral bodily constraints

Understanding rhythmic multimodal communication also requires a still underdeveloped understanding of peripheral bodily constraints (henceforth biomechanics) in the production of multimodal signals. Here we overview findings which show how multimodal signaling sometimes exploits physical properties of the body in the construction of temporally complex signals.

Speech is putatively a superordinate mode of coordination between what were originally stable independent vocal and mandibular action routines [47]. In chimpanzees (*Pan Troglodytes*), non-vocal lip smacking occurs in the theta range (~3-8Hz) typical of the speech envelope and labial kinematics of human speech [48]. Marmosets (*Callithrix jacchus*) occupy bistable modes of vocal-articulatory coordination, where mandibular oscillation is only synchronized at the characteristic theta range with vocal modulations at the final but not starting segments of the call [49]. Similarly, in the zebra finch (*Taeniopygia guttata*), respiratory pulses are timed with syrinx activity and rapid beak movements, the coordination of which is held to sustain the highly varied vocalization repertoire of this bird species [50]. Human speech is characterized by even more hierarchically nested levels of such coordinated periodicities of effectors and is in this sense multimodal [51] .

Human communicative hand gestures have acceleration peaks co-occurrent with emphatic stress in speech, which are tightly and dynamically coupled under adverse conditions, though with more temporal variability for more complex symbolizing gestures [52]. This coupling of gestures' acceleration-induced forces and speech can arise

biomechanically from upper limb-respiratory coupling, e.g., by soliciting anticipatory muscle adjustments to stabilize posture during gesture [53], which also include respiratory-controlling muscles supporting speech-vocalization [54]. Comparable biomechanical interactions and synergies have been found in other animals long before such associations were raised to explain aspects of human multimodal prosody. In brown-headed cowbirds (*Molothrus ater*) vocalizations are produced with specific respiratory-related abdominal muscle activity. Such modulations are reduced during vocalizing while moving the wings for visual displaying, even though air sac pressure is maintained. This suggests that visual displays in cowbirds biomechanically interact with respiratory dynamics supporting vocalization [55]. During their more vigorous wing-displays, these birds are vocally silent, likely so as to avoid biomechanical instability of singing and moving vigorously at the same time. Such biomechanical interactions are consistent with findings of the wing-beats of flying bats (e.g., *Pteronotus parnellii*), which are synchronized with echo-vocalizations due to locomotion-respiratory biomechanical synergies [56]. The echo-vocalizations during flight are often isochronously structured (at 6-12 Hz), and this rhythmic ability is attributed to locomotion-respiratory couplings as they share a temporal structure. However, isochrony (at 12-24Hz) has also been observed in stationary bats when producing social vocalizations [57]. In this way, biomechanical stabilities from one domain may have scaffolded the rhythmic vocal capabilities that are sustained in social vocal domains [58].

Rhesus macaques assume different facial postures with particular vocalizations. Lips usually protrude when emitting coos or grunts (e.g., during mother-infant contact or group progression). During the emission of screams (e.g., copulation or threats), lips retract [59]. In macaques, facial gestures are associated with peculiar vocal tract shapes, which influence acoustic signals during phonation [60] and can be discriminated by conspecific listeners [61]. Relatedly, in humans, *perceiving* lip postures allows the perceiver to derive a /ba/ or /pa/ from an auditory signal. It is the auditory-visual-motor co-regularity that makes visual or haptic perception of articulatory gestures possible in this classic McGurk-effect [62]. Recently a "manual gesture McGurk-effect" has been discovered [63]. When asked to detect a particular lexical stress in a uniformly stressed speech sequence, participants who *see* a hand gesture's beat timed with a particular speech segment tend to *hear* a lexical stress for that segment [63]. We think it is possible that the gesture-speech-respiratory link as reviewed above, is actually important for understanding the manual McGurk-effect as listeners attune to features of the visual-acoustic signal that are informative about such coordinated modes of production [64]. Similarly, communicative gestures can also influence the heard duration of musical notes. For example, the absolute duration of a percussive tone sounds longer to an audience member when seeing a long- vs short-percussion gesture [65,66].

Furthermore, spontaneous movements are naturally elicited by music. Whether this spontaneous movement stems from generalizable cross-modal associations is debated, but they might be identified when properly related to biomechanics. For instance, hierarchical bodily representations of meter can be elicited in spontaneous music-induced movement, with different aspects of the meter embodied in hand, torso, or full arm movements [67]. Additionally, specific coordination patterns emerge between different body parts of interacting musicians during musical improvisation [68]. Thus what one hears in music might be constrained to what body part can be optimally temporally aligned with a feature in the music.

To detect multimodal cues in this way may be very closely related to indexical signals, such as hearing the potential strength of a conspecific from vocal qualities [61,69]. Indexical signals are often the result of perceptual and/or morphological specialization to detect/convey features from multimodal couplings. For example, frogs (*Physalaemus pustulosus*) and frog-eating bats (*Trachops cirrhosus*) have learned to attune to frog calls *in relation* to the water ripples produced by the calling frog's vocal sac deformations [70]. Similarly, crested pigeons (*Ochyphaps lophotes*) are alarmed by the sounds of high velocity wing beats of conspecifics, where the feathers turn out to have morphologically evolved to produce the aeroelastic flutter needed to sustain these unique alarm calls during fleeing locomotion [55]. In broad-tailed hummingbirds (*Selasphorus platycercus*) the characteristic high-speed courtship dives seem to be driven to attain exactly the right speeds to elicit sonification from aeroelastic flutter, which is synchronized with attaining the correct angle transition relative to the to-be-impressed perceiver so that the gorget dramatically changes color during sound production [72]. In sum, multimodal communication sometimes involves a specialized exploitation or attunement of physics that constrains particular (combined) modes of acting (with the environment).

Note that the multimodal information embedded in communicative acoustic signals can have impacts on complex communication in humans too. Speakers who cannot see but only hear each other tend to align patterns of postural sway suggesting that vocal features are used to coordinate a wider embodied context [73]. These emergent coordinations are found to increase social affiliation and can align bodily processes [74]. For example, synchronized drumming in groups synchronizes physiology, aligning participants' heartbeats [75]. Further, visual observation of interpersonal synchronous movement between others may lead observers to rate higher levels of rapport (liking) between the interacting individuals [76], and increase an audience's affective and aesthetic enjoyment of group dance performance [77].

To conclude, we have overviewed examples of peripheral bodily constraints which influence the perception and production of multimodal signals across species. Specifically, these biomechanical processes mediate the temporal structuring of multimodal communicative signals.

Section 4. Social Level: Complex rhythms in interactive multimodal communication

In this section we overview how social interaction complexifies the rhythms which are sustained in communication relative to the rhythms that would arise out of more simple sending or receiving of signals.

Temporal structure is often rhythmic, but studies have also found quasi-rhythmic structure in sounds of speech, music, and animal communication [78], and likewise for movements produced while talking, singing, or performing music [79]. The multiscale character of these sounds and movements is readily illustrated in speech—phonemes of varying durations combine to create longer syllabic units with more variability in duration, which combine to form phrases with even more variability in length, and so on, thus creating quasi-rhythmicity at each timescale.

The durations of linguistic units like phonemes and syllables are difficult to measure in the acoustic speech signal, but they generally correspond to modulations in a specific feature of the acoustic signal, called the amplitude envelope. Within the amplitude envelope, units are

expressed in terms of bursts and lulls of energy, and their temporal patterning can be distilled in the timing of bursts via peak amplitudes. Speech analysis [80] shows that smaller bursts cluster to form larger bursts, where larger bursts cluster to form even larger bursts across timescales that roughly correspond with (phonemic, syllabic, phrasal) units of language. Musical recordings also exhibit degrees of multiscale structure whose specifics depend on the genre of music or type of speech performance [78]. Even recordings of animal vocalizations have been found to exhibit multiscale structure using those same analysis methods. While we do not have access to the underlying units, recordings of communicative vocalizations produced by killer whales were found to have a quasi-rhythmic structure across timescales surprisingly similar to human speech interactions [78].

Multiscale structure in speech and music is also multimodal. Analyses of sounds and movements in video recordings have found coordinated multiscale structures in the amplitudes of co-speech face, head, and body movements [79], and the degree of coordination in speech sounds and movements depends on the communicative context. Studies of rhythmic structure have also found that visual communicative signals are tightly coordinated with the acoustic signals of speech [3]. However, while gestures with a beating quality coincide closely with pitch peaks, on the semantic level object- or action-depicting gestures frequently precede corresponding lexical items by several hundred milliseconds [81]. Facial signals, too, can precede the speech they relate to [82]. Variable timing is most obvious if we consider multimodal utterances in their entirety, where speech is embedded in a rich infrastructure of visual signals coming from the hands, head, face, torso, etc. [3]. These different signals are typically not aligned in time but distributed over the entire length of utterances and beyond, with varying onsets and offsets.

Typically, multimodal utterances in human social interaction are produced within a scaffold of speaking turns. Sacks et al. [83] propose that interlocutors abide by a clear set of rules which, combined with linguistic information (semantics, pragmatics, prosody and syntax), afford precise timing of turns, yielding minimal gaps and overlaps. Indeed, quantitative cross-language analyses support this tight temporal coupling [84], in line with a putative “interaction engine” providing cognitive-interactional predispositions for this human ability [85], though gestural turn exchanges in bonobos point towards an evolutionary precursor [86].

Rhythmical structure may further facilitate the temporal coupling of turns. Wilson and Wilson [87] specify a mechanism by which interlocutors’ endogenous oscillators are anti-phase coupled, allowing next speakers to launch their turn ‘on time’, while decreasing the chance of overlap. This may be enhanced through temporal projections derived from linguistic information [88], but the rhythmical abilities grounding this mechanism are evolutionarily basic [89]. Wild nonhuman primates, like indris, gibbons and chimpanzees, show coordination during joint vocal output, suggesting the ability to coordinate to auditory rhythms [90,91]. The captive chimpanzee Ai was able to synchronize her keyboard tapping with an acoustic stimulus [92], and captive macaques can flexibly adjust their tapping in anticipation of the beat of a visual metronome [34]. Moreover, cotton-top tamarins and marmosets have been observed to avoid initiating and adjust the duration and onset of their calls such that they avoid interfering noise [93].

However, conversational turn-taking is also characterized by temporal variation, including periods of overlap and gaps ranging up to hundreds of milliseconds [84,94]. The full breadth of factors influencing turn transition times remains opaque, but turn duration, syntactic complexity, word frequency, and social action are some of them [95]. A coupled-oscillator turn-taking mechanism can accommodate this large variation in turn timing, since entrained interlocutors could begin speaking at any new anti-phased periodic cycle [87,89]. A recent study based on telephone interactions shows a quasi-rhythmic structure regulated by turn-by-turn negative autocorrelations [96]. The coupled-oscillator mechanism that may form the basis for dealing with quasi-rhythmicity at the interactional level may also govern communication in non-human species, such as the interactional synchronization of non-isochronous call patterns in the katydid species *Mecopoda* [97].

To conclude, the temporal organization of intentional communication is an intricate matter, characterized, on one hand, by synchrony serving the amplification of signals or specific features/components thereof, as well as semantic enhancement and smooth coordination between interlocutors. On the other hand, the temporal organization is characterized by quasi-rhythmic, multiscale structure within and across modalities, serving complex communication and coordination patterns that are widespread in communicative animal vocalizations, human speech, and even music.

Conclusion

We have argued that to understand communicative rhythms which characterize animal communication, a multimodal perspective is necessary and multiple levels need to be examined. The current overview takes a first step towards a multilevel multimodal approach, showing how each level (neural, bodily, interactive) uniquely contributes to the communicative rhythms of animals. We think that when processes on these levels are understood we can come to understand why the rhythms of for example human conversation are so complexly varied. Though we have addressed the unique contributions at each level independently, the biggest challenge is understanding how levels intersect.

A historic lesson in this regard comes from early theories about human vocalization. Early theories held that phonation was actively neurally driven, such that active muscle contractions would be needed to complete each vocal fold cycle [98]. This hypothesis was soon refuted in favor of a biomechanical theory [99], which correctly posited that vocal fold oscillation arises out of more neurally passive dynamics. Namely, vocal fold oscillations arise due to air pressure flux around a tensed elastic material (i.e., vocal folds). Similarly, neurally passive dynamics have been discovered in subsonic phonations in elephant (*Loxodonta africana*) trunks [100]. But interestingly, it turns out that for several cat species low frequency purring *is* actively neuro-muscularly driven to complete a cycle [101]. The lesson is that the neural-cognitive mechanisms that are invoked in our explanations of rhythmic communication will crucially depend on our knowledge of biomechanics, and any redundancies present biomechanically can completely reshape the type of neural-cognitive control mechanisms that need be invoked. In the same way, understanding the unique neural constraints can lead to the discovery that neural-cognitive mechanisms need to be in place to exploit certain bodily capacities [102]. A recent integrative approach has been proposed in the understanding of beat perception and motor synchronization, where it is suggested that a *network of biological*

oscillators are at play when moving to a rhythm, which involves more neurally passive dynamic biomechanics and neural processes [28]. Finally, social interactions allow for new rhythmic stabilities that are simply absent or qualitatively different in nature than non-interactive setups [103]. Indeed, there are increasingly louder calls for action for understanding neural processes as sometimes softly assembling into a wider distributed multi-person system in social interactions [104–106]. The current contribution further underlines a call for such a multiscale investigation of temporal rhythms of multimodal communication, where neural processes are properly embedded in bodily processes unfolding in social interaction.

References

1. Partan SR, Marler P. 2005 Issues in the classification of multimodal communication signals. *Am. Nat.* **166**, 231–245. (doi:10.1086/431246)
2. Fröhlich M, Sievers C, Townsend SW, Gruber T, Schaik CP van. 2019 Multimodal communication and language origins: integrating gestures and vocalizations. *Biol. Rev.* **94**, 1809–1829. (doi:10.1111/brv.12535)
3. Holler J, Levinson SC. 2019 Multimodal language processing in human communication. *Trends Cogn. Sci.* **23**, 639–652. (doi:10.1016/j.tics.2019.05.006)
4. Stein BE, Stanford TR. 2008 Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* **9**, 255–266. (doi:10.1038/nrn2331)
5. Hebets EA, Papaj DR. 2005 Complex signal function: developing a framework of testable hypotheses. *Behav. Ecol. Sociobiol.* **57**, 197–214. (doi:10.1007/s00265-004-0865-7)
6. Dimensions.Guide | Database of Dimensioned Drawings. See https://www.dimensions.guide (accessed on 1 May 2019).
7. Chilton J. 2020 Brain outline. (doi:10.5281/zenodo.3925989)
8. Partan SR, Marler P. 1999 Communication Goes Multimodal. *Science* **283**, 1272–1273. (doi:10.1126/science.283.5406.1272)
9. Sandler W. 2018 The Body as Evidence for the Nature of Language. *Front. Psychol.* **9**. (doi:10.3389/fpsyg.2018.01782)
10. Halfwerk W, Varkevisser J, Simon R, Mendoza E, Scharff C, Riebel K. 2019 Toward Testing for Multimodal Perception of Mating Signals. *Front. Ecol. Evol.* **7**. (doi:10.3389/fevo.2019.00124)
11. Stein BE, Meredith MA. 1993 *The merging of the senses*. Massachusetts: MIT Press.
12. Xu J, Yu L, Stanford TR, Rowland BA, Stein BE. 2015 What does a neuron learn from multisensory experience? *J. Neurophysiol.* **113**, 883–889. (doi:10.1152/jn.00284.2014)
13. Yu L, Cuppini C, Xu J, Rowland BA, Stein BE. 2019 Cross-modal competition: The default computation for multisensory processing. *J. Neurosci.* **39**, 1374–1385. (doi:10.1523/JNEUROSCI.1806-18.2018)
14. Meredith MA, Stein BE. 1986 Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. Neurophysiol.* **56**, 640–662. (doi:10.1152/jn.1986.56.3.640)
15. Stein BE, Meredith MA, Huneycutt WS, McDade L. 1989 Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. *J. Cogn. Neurosci.* **1**, 12–24. (doi:10.1162/jocn.1989.1.1.12)
16. Burnett LR, Stein BE, Perrault TJ, Wallace MT. 2007 Excitotoxic lesions of the superior colliculus preferentially impact multisensory neurons and multisensory

integration. *Exp. Brain Res.* **179**, 325–338. (doi:10.1007/s00221-006-0789-8)
17. Khandhadia AP, Murphy AP, Romanski LM, Bizley JK, Leopold DA. 2021 Audiovisual integration in macaque face patch neurons. *Curr. Biol.* (doi:10.1016/j.cub.2021.01.102)
18. Narins PM, Grabul DS, Soma KK, Gaucher P, Hödl W. 2005 Cross-modal integration in a dart-poison frog. *Proc. Natl. Acad. Sci.* **102**, 2425–2429. (doi:10.1073/pnas.0406407102)
19. Uy JAC, Safran RJ. 2013 Variation in the temporal and spatial use of signals and its implications for multimodal communication. *Behav. Ecol. Sociobiol.* **67**, 1499–1511. (doi:10.1007/s00265-013-1492-y)
20. Senkowski D, Schneider TR, Foxe JJ, Engel AK. 2008 Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci.* **31**, 401–409. (doi:10.1016/j.tins.2008.05.002)
21. Bauer A-KR, Debener S, Nobre AC. 2020 Synchronisation of neural oscillations and cross-modal influences. *Trends Cogn. Sci.* **24**, 481–495. (doi:10.1016/j.tics.2020.03.003)
22. Drijvers L, Spaak E, Jensen O. 2020 Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual semantic information. *bioRxiv* , 2020.04.29.067454. (doi:10.1101/2020.04.29.067454)
23. Zhigalov A, Herring JD, Herpers J, Bergmann TO, Jensen O. 2019 Probing cortical excitability using rapid frequency tagging. *NeuroImage* **195**, 59–66. (doi:10.1016/j.neuroimage.2019.03.056)
24. Duecker K, Gutteling TP, Herrmann CS, Jensen O. 2020 No evidence for entrainment: endogenous gamma oscillations and rhythmic flicker responses coexist in visual cortex. *bioRxiv* , 2020.09.02.279497. (doi:10.1101/2020.09.02.279497)
25. Eitan Z, Granot RY. 2006 Musical Parameters and Listeners Images of Motion: : Musical Parameters and Listeners Images of Motion. *Music Percept. Interdiscip. J.* **23**, 221–248. (doi:10.1525/mp.2006.23.3.221)
26. Küssner MB, Tidhar D, Prior HM, Leech-Wilkinson D. 2014 Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Front. Psychol.* **5**. (doi:10.3389/fpsyg.2014.00789)
27. Schaefer RS, Beijer LJ, Seuskens W, Rietveld TCM, Sadakata M. 2016 Intuitive visualizations of pitch and loudness in speech. *Psychon. Bull. Rev.* **23**, 548–555. (doi:10.3758/s13423-015-0934-0)
28. Damm L, Varoqui D, De Cock VC, Dalla Bella S, Bardy B. 2020 Why do we move to the beat? A multi-scale approach, from physical principles to brain dynamics. *Neurosci. Biobehav. Rev.* **112**, 553–584. (doi:10.1016/j.neubiorev.2019.12.024)
29. Morillon B, Baillet S. 2017 Motor origin of temporal predictions in auditory attention. *Proc. Natl. Acad. Sci.* **114**, E8913–E8921. (doi:10.1073/pnas.1705373114)
30. Schubotz RI, Friederici AD, von Cramon DY. 2000 Time perception and motor timing: a common cortical and subcortical basis revealed by fMRI. *NeuroImage* **11**, 1–12. (doi:10.1006/nimg.1999.0514)
31. Repp BH, Penel A. 2004 Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychol. Res.* **68**, 252–270. (doi:10.1007/s00426-003-0143-8)
32. Hove MJ, Fairhurst MT, Kotz SA, Keller PE. 2013 Synchronizing with auditory and visual rhythms: An fMRI assessment of modality differences and modality

appropriateness. *NeuroImage* **67**, 313–321. (doi:10.1016/j.neuroimage.2012.11.032)
33. Peña M, Langus A, Gutiérrez C, Huepe-Artigas D, Nespor M. 2016 Rhythm on Your Lips. *Front. Psychol.* **7**. (doi:10.3389/fpsyg.2016.01708)
34. Gámez J, Yc K, Ayala YA, Dotov D, Prado L, Merchant H. 2018 Predictive rhythmic tapping to isochronous and tempo changing metronomes in the nonhuman primate. *Ann. N. Y. Acad. Sci.* (doi:10.1111/nyas.13671)
35. Merchant H, Honing H. 2014 Are non-human primates capable of rhythmic entrainment? Evidence for the gradual audiomotor evolution hypothesis. *Front. Neurosci.* **7**. (doi:10.3389/fnins.2013.00274)
36. Drijvers L, Özyürek A, Jensen O. 2018 Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech. *J. Cogn. Neurosci.* **30**, 1086–1097. (doi:10.1162/jocn_a_01301)
37. Pearce MT, Ruiz MH, Kapasi S, Wiggins GA, Bhattacharya J. 2010 Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage* **50**, 302–313. (doi:10.1016/j.neuroimage.2009.12.019)
38. Repp BH. 2005 Sensorimotor synchronization: A review of the tapping literature. *Psychon. Bull. Rev.* **12**, 969–992. (doi:10.3758/BF03206433)
39. Roy C, Lagarde J, Dotov D, Dalla Bella S. 2017 Walking to a multisensory beat. *Brain Cogn.* **113**, 172–183. (doi:10.1016/j.bandc.2017.02.002)
40. Escoffier N, Herrmann CS, Schirmer A. 2015 Auditory rhythms entrain visual processes in the human brain: Evidence from evoked oscillations and event-related potentials. *NeuroImage* **111**, 267–276. (doi:10.1016/j.neuroimage.2015.02.024)
41. Brown S, Martinez MJ, Parsons LM. 2006 The neural basis of human dance. *Cereb. Cortex N. Y. N 1991* **16**, 1157–1167. (doi:10.1093/cercor/bhj057)
42. Schaefer RS, Morcom AM, Roberts N, Overy K. 2014 Moving to music: Effects of heard and imagined musical cues on movement-related brain activity. *Front. Hum. Neurosci.* **8**. (doi:10.3389/fnhum.2014.00774)
43. Brown RM, Penhune VB. 2018 Efficacy of auditory versus motor learning for skilled and novice performers. *J. Cogn. Neurosci.* **30**, 1657–1682. (doi:10.1162/jocn_a_01309)
44. van Vugt FT, Kafczyk T, Kuhn W, Rollnik JD, Tillmann B, Altenmüller E. 2016 The role of auditory feedback in music-supported stroke rehabilitation: A single-blinded randomised controlled intervention. *Restor. Neurol. Neurosci.* **34**, 297–311. (doi:10.3233/RNN-150588)
45. Moore E, Schaefer RS, Bastin ME, Roberts N, Overy K. 2017 Diffusion tensor MRI tractography reveals increased fractional anisotropy (FA) in arcuate fasciculus following music-cued motor training. *Brain Cogn.* **116**, 40–46. (doi:10.1016/j.bandc.2017.05.001)
46. Schaefer RS. 2014 Auditory rhythmic cueing in movement rehabilitation: findings and possible mechanisms. *Philos. Trans. R. Soc. B Biol. Sci.* **369**. (doi:10.1098/rstb.2013.0402)
47. MacNeilage PF. 1998 The frame/content theory of evolution of speech production. *Behav. Brain Sci.* **21**, 499–511. (doi:10.1017/S0140525X98001265)
48. Pereira AS, Kavanagh E, Hobaiter C, Slocombe KE, Lameira AR. 2020 Chimpanzee lip-smacks confirm primate continuity for speech-rhythm evolution. *Biol. Lett.* **16**, 20200232. (doi:10.1098/rsbl.2020.0232)
49. Risueno-Segovia C, Hage SR. 2020 Theta synchronization of phonatory and

articulatory systems in marmoset monkey vocal production. *Curr. Biol.* (doi:10.1016/j.cub.2020.08.019)
50. Goller F, Cooper BG. 2004 Peripheral motor dynamics of song production in the zebra finch. *Ann. N. Y. Acad. Sci.* **1016**, 130–152. (doi:10.1196/annals.1298.009)
51. Kelso JAS, Tuller B, Harris K. 1983 A "Dynamic Pattern" perspective on the control and coordination of movement. In *The production of speech*, Berlin: Springer-Verlag.
52. Pouw W, Dixon JA. 2019 Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cogn. Sci.* **43**, e12721. (doi:10.1111/cogs.12721)
53. Hodges PW, Richardson CA. 1997 Feedforward contraction of transversus abdominis is not influenced by the direction of arm movement. *Exp. Brain Res.* **114**, 362–370. (doi:10.1007/pl00005644)
54. Pouw W, Harrison SJ, Esteve-Gibert N, Dixon JA. 2020 Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *J. Acoust. Soc. Am.* **148**, 1231–1247. (doi:10.1121/10.0001730)
55. Cooper BG, Goller F. 2004 Multimodal signals: enhancement and constraint of song motor patterns by visual display. *Science* **303**, 544–546. (doi:10.1126/science.1091099)
56. Lancaster WC, Henson OW, Keating AW. 1995 Respiratory muscle activity in relation to vocalization in flying bats. *J. Exp. Biol.* **198**, 175–191.
57. Burchardt LS, Norton P, Behr O, Scharff C, Knörnschild M. 2019 General isochronous rhythm in echolocation calls and social vocalizations of the bat *Saccopteryx bilineata*. *R. Soc. Open Sci.* **6**, 181076. (doi:10.1098/rsos.181076)
58. Larsson M, Richter J, Ravignani A. 2019 Bipedal steps in the development of rhythmic behavior in humans. *Music Sci.* **2**, 2059204319892617. (doi:10.1177/2059204319892617)
59. Hauser MD, Evans CS, Marler P. 1993 The role of articulation in the production of rhesus monkey, Macaca mulatta, vocalizations. *Anim. Behav.* **45**, 423–433. (doi:10.1006/anbe.1993.1054)
60. Fitch WT. 1997 Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J. Acoust. Soc. Am.* **102**, 1213–1222. (doi:10.1121/1.421048)
61. Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. 2007 Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr. Biol.* **17**, 425–430. (doi:10.1016/j.cub.2007.01.029)
62. Fowler CA, Dekle DJ. 1991 Listening with eye and hand: Cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 816–828. (doi:10.1037/0096-1523.17.3.816)
63. Bosker HR, Peeters D. 2020 Beat gestures influence which speech sounds you hear. *Proc. R. Soc. B Biol. Sci.* (doi:10.1101/2020.07.13.200543)
64. Pouw W, Paxton A, Harrison SJ, Dixon JA. 2020 Acoustic information about upper limb movement in voicing. *Proc. Natl. Acad. Sci.* **117**, 11364–11367. (doi:10.1073/pnas.2004163117)
65. Schutz M, Lipscomb S. 2016 Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception* (doi:10.1068/p5635)
66. Cai ZG, Connell L, Holler J. 2013 Time does not flow without language: Spatial distance affects temporal duration regardless of movement or direction. *Psychon. Bull. Rev.* **20**, 973–980. (doi:10.3758/s13423-013-0414-3)
67. Toiviainen P, Luck G, Thompson MR. 2010 Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Percept.* **28**, 59–70.

(doi:10.1525/mp.2010.28.1.59)
68. Walton AE, Washburn A, Langland-Hassan P, Chemero A, Kloos H, Richardson MJ. 2018 Creating time: Social collaboration in music improvisation. *Top. Cogn. Sci.* **10**, 95–119. (doi:10.1111/tops.12306)
69. Pisanski K, Cartei V, McGettigan C, Raine J, Reby D. 2016 Voice modulation: A window into the origins of human vocal control? *Trends Cogn. Sci.* **20**, 304–318. (doi:10.1016/j.tics.2016.01.002)
70. Halfwerk W, Jones PL, Taylor RC, Ryan MJ, Page RA. 2014 Risky Ripples Allow Bats and Frogs to Eavesdrop on a Multisensory Sexual Display. *Science* **343**, 413–416. (doi:10.1126/science.1244812)
71. Murray TG, Zeil J, Magrath RD. 2017 Sounds of Modified Flight Feathers Reliably Signal Danger in a Pigeon. *Curr. Biol.* **27**, 3520-3525.e4. (doi:10.1016/j.cub.2017.09.068)
72. Hogan BG, Stoddard MC. 2018 Synchronization of speed, sound and iridescent color in a hummingbird aerial courtship dive. *Nat. Commun.* **9**, 5260. (doi:10.1038/s41467-018-07562-7)
73. Shockley K, Baker AA, Richardson MJ, Fowler CA. 2007 Articulatory constraints on interpersonal postural coordination. *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 201–208. (doi:10.1037/0096-1523.33.1.201)
74. Wiltermuth SS, Heath C. 2009 Synchrony and cooperation. *Psychol. Sci.* **20**, 1–5. (doi:10.1111/j.1467-9280.2008.02253.x)
75. Gordon I, Gilboa A, Cohen S, Milstein N, Haimovich N, Pinhasi S, Siegman S. 2020 Physiological and behavioral synchrony predict group cohesion and performance. *Sci. Rep.* **10**, 8484. (doi:10.1038/s41598-020-65670-1)
76. McEllin L, Knoblich G, Sebanz N. 2020 Synchronicities that shape the perception of joint action. *Sci. Rep.* **10**, 15554. (doi:10.1038/s41598-020-72729-6)
77. Vicary S, Sperling M, Zimmermann J von, Richardson DC, Orgs G. 2017 Joint action aesthetics. *PLOS ONE* **12**, e0180101. (doi:10.1371/journal.pone.0180101)
78. Kello CT, Bella SD, Médé B, Balasubramaniam R. 2017 Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *J. R. Soc. Interface* **14**, 20170231. (doi:10.1098/rsif.2017.0231)
79. Alviar C, Dale R, Dewitt A, Kello C. 2020 Multimodal Coordination of Sound and Movement in Music and Speech. *Discourse Process.* **57**, 682–702. (doi:10.1080/0163853X.2020.1768500)
80. Falk S, Kello CT. 2017 Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* **163**, 80–86. (doi:10.1016/j.cognition.2017.02.017)
81. Bekke M ter, Drijvers L, Holler J. 2020 The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. (doi:10.31234/osf.io/b5zq7)
82. Kaukomaa T, Peräkylä A, Ruusuvuori J. 2013 Turn-opening smiles: Facial expression constructing emotional transition in conversation. *J. Pragmat.* **55**, 21–42. (doi:10.1016/j.pragma.2013.05.006)
83. Sacks H, Schegloff EA, Jefferson G. 1974 A Simplest systematics for the organization of turn-taking for conversation. *Language* **50**, 696–735.
84. Stivers T *et al.* 2009 Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci.* **106**, 10587–10592. (doi:10.1073/pnas.0903616106)

85. Levinson SC. 2019 Interactional foundations of language: The interaction engine hypothesis. In *Human language: From genes and brain to behavior*, pp. 189–200. MIT Press.
86. Fröhlich M, Kuchenbuch P, Müller G, Fruth B, Furuichi T, Wittig RM, Pika S. 2016 Unpeeling the layers of language: Bonobos and chimpanzees engage in cooperative turn-taking sequences. *Sci. Rep.* **6**, 25887. (doi:10.1038/srep25887)
87. Wilson M, Wilson TP. 2005 An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* **12**, 957–968. (doi:10.3758/BF03206432)
88. Levinson SC. 2016 Turn-taking in human communication--Origins and implications for language processing. *Trends Cogn. Sci.* **20**, 6–14. (doi:10.1016/j.tics.2015.10.010)
89. Takahashi DY, Narayanan DZ, Ghazanfar AA. 2013 Coupled oscillator dynamics of vocal turn-taking in monkeys. *Curr. Biol.* **23**, 2162–2168. (doi:10.1016/j.cub.2013.09.005)
90. Geissmann T, Orgeldinger M. 2000 The relationship between duet songs and pair bonds in siamangs, Hylobates syndactylus. *Anim. Behav.* **60**, 805–809. (doi:10.1006/anbe.2000.1540)
91. Gamba M, Torti V, Estienne V, Randrianarison RM, Valente D, Rovara P, Bonadonna G, Friard O, Giacoma C. 2016 The Indris have got rhythm! Timing and pitch variation of a primate song examined between sexes and age classes. *Front. Neurosci.* **10**. (doi:10.3389/fnins.2016.00249)
92. Hattori Y, Tomonaga M, Matsuzawa T. 2013 Spontaneous synchronized tapping to an auditory rhythm in a chimpanzee. *Sci. Rep.* **3**, 1566. (doi:10.1038/srep01566)
93. Egnor SER, Wickelgren JG, Hauser MD. 2007 Tracking silence: adjusting vocal production to avoid acoustic interference. *J. Comp. Physiol. A* **193**, 477–483. (doi:10.1007/s00359-006-0205-7)
94. Holler J, Kendrick KH, Levinson SC. 2018 Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychon. Bull. Rev.* **25**, 1900–1908. (doi:10.3758/s13423-017-1363-z)
95. Roberts SG, Torreira F, Levinson SC. 2015 The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Front. Psychol.* **6**. (doi:10.3389/fpsyg.2015.00509)
96. Pouw W, Holler J. 2020 Timing in conversation is dynamically adjusted turn by turn: Evidence for lag-1 negatively autocorrelated turn taking times in telephone conversation. (doi:10.31234/osf.io/b98da)
97. Nityananda V, Balakrishnan R. 2020 Synchrony of complex signals in an acoustically communicating katydid. *J. Exp. Biol.* (doi: 10.1242/jeb.241877)
98. Husson R, Garde ÉJ, Richard A. 1952 L'acoustique des salles du point de vue du chanteur et de l'orateur. *Ann. Télécommunications* **7**, 58–74. (doi:10.1007/BF03017103)
99. Titze IR. 1973 The Human Vocal Cords: A Mathematical Model. *Phonetica* **28**, 129–170. (doi:10.1159/000259453)
100. Herbst CT, Stoeger AS, Frey R, Lohscheller J, Titze IR, Gumpenberger M, Fitch WT. 2012 How low can you go? Physical production mechanism of elephant infrasonic vocalizations. *Science* **337**, 595–599. (doi:10.1126/science.1219712)
101. Peters G. 2002 Purring and similar vocalizations in mammals. *Mammal Rev.* **32**, 245–271. (doi:https://doi.org/10.1046/j.1365-2907.2002.00113.x)
102. Fitch WT. 2018 The Biology and Evolution of Speech: A Comparative Analysis. *Annu. Rev. Linguist.* **4**, 255–279. (doi:10.1146/annurev-linguistics-011817-045748)

103. Tognoli E, Zhang M, Fuchs A, Beetle C, Kelso JAS. 2020 Coordination dynamics: A foundation for understanding social behavior. *Front. Hum. Neurosci.* **14**. (doi:10.3389/fnhum.2020.00317)
104. De Jaegher H, Di Paolo E, Gallagher S. 2010 Can social interaction constitute social cognition? *Trends Cogn. Sci.* **14**, 441–447. (doi:10.1016/j.tics.2010.06.009)
105. Dumas G, Kelso JAS, Nadel J. 2014 Tackling the social cognition paradox through multi-scale approaches. *Front. Psychol.* **5**. (doi:10.3389/fpsyg.2014.00882)
106. Redcay E, Schilbach L. 2019 Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat. Rev. Neurosci.* **20**, 495–505. (doi:10.1038/s41583-019-0179-4)